

SPRINGONE2GX

WASHINGTON, DC

Who Needs Batch?

By Michael Minella and Gunnar Hillert

@michaelminella, @ghillert





springone **2GX**

Gunnar Hillert

Twitter: @ghillert

Website: <http://blog.hillert.com>

Conference: <http://devnexus.com>

Michael Minella

Twitter: @michaelminella

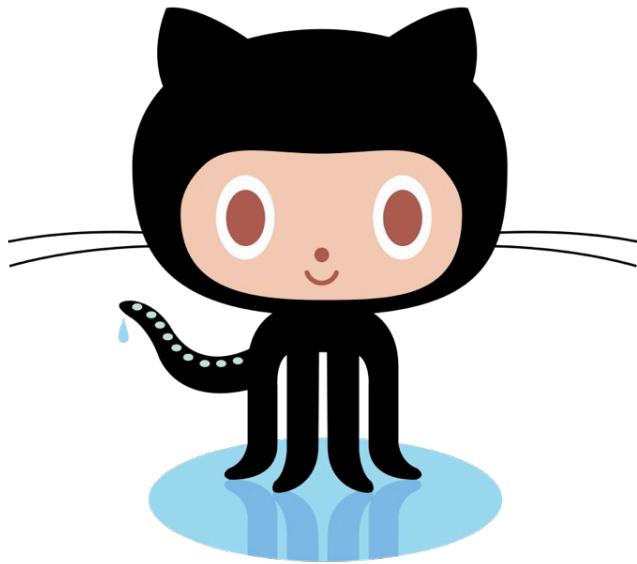
Podcast: <http://javaOffHeap.com>

or @OffHeap

Website: <http://spring.io>



springone 2GX



<https://github.com/ghillert/s1-2gx2015-who-needs-batch>



**WE HAVE AN
ANNOUNCEMENT**



WE HAVE A DISEASE



OBJECTIVIUS SHINIUM SYNDROMUS

A Gollum-like creature, with a pale, gaunt face and large, bulbous nose, is shown from the waist up, crouching on a dark, craggy rock formation. It has long, thin fingers and toes. In the background, several tall, vertical, translucent crystals with a metallic sheen and a warm, golden glow are visible against a dark, hazy sky.

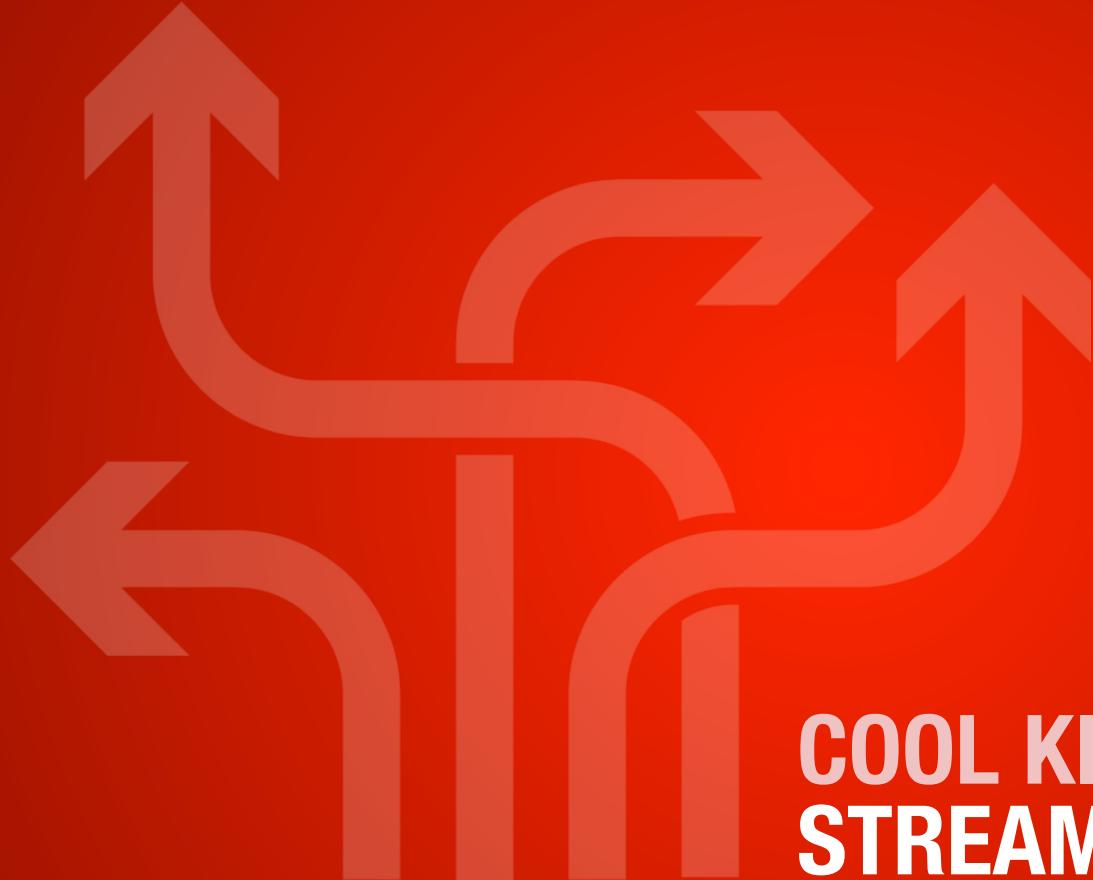
SHINY OBJECT SYNDROME



IT MAKES US
INNOVATE

ROAD TO RECOVERY





**COOL KIDS ARE DOING
STREAM PROCESSING**



Google Cloud Dataflow



Tigon



STORM



samza



springone 2GX

Say goodbye to batch: Big Data

strataconf.com/big-data-conference-uk-2015/public/schedule/detail/39966

PRESENTED BY O'REILLY AND CLOUDERA SAN JOSE • LONDON • NEW YORK • SINGAPORE

Strata+Hadoop WORLD

Make Data Work

5-7 May, 2015 • London, UK

SCHEDULE | SPEAKERS | SPONSOR PAVILION | RESOURCES | VENUE | ABOUT | ACCOUNT

[Join Attendee Network](#)

[Add to Your Schedule](#)

Say goodbye to batch

Tyler Akidau (Google)
16:15-16:55 Thursday, 7/05/2015
Hadoop & Beyond
Location: Buckingham Room - Palace Suite
Average rating: ★★★★☆ (4.14, 7 ratings)

Slides: [external link](#)

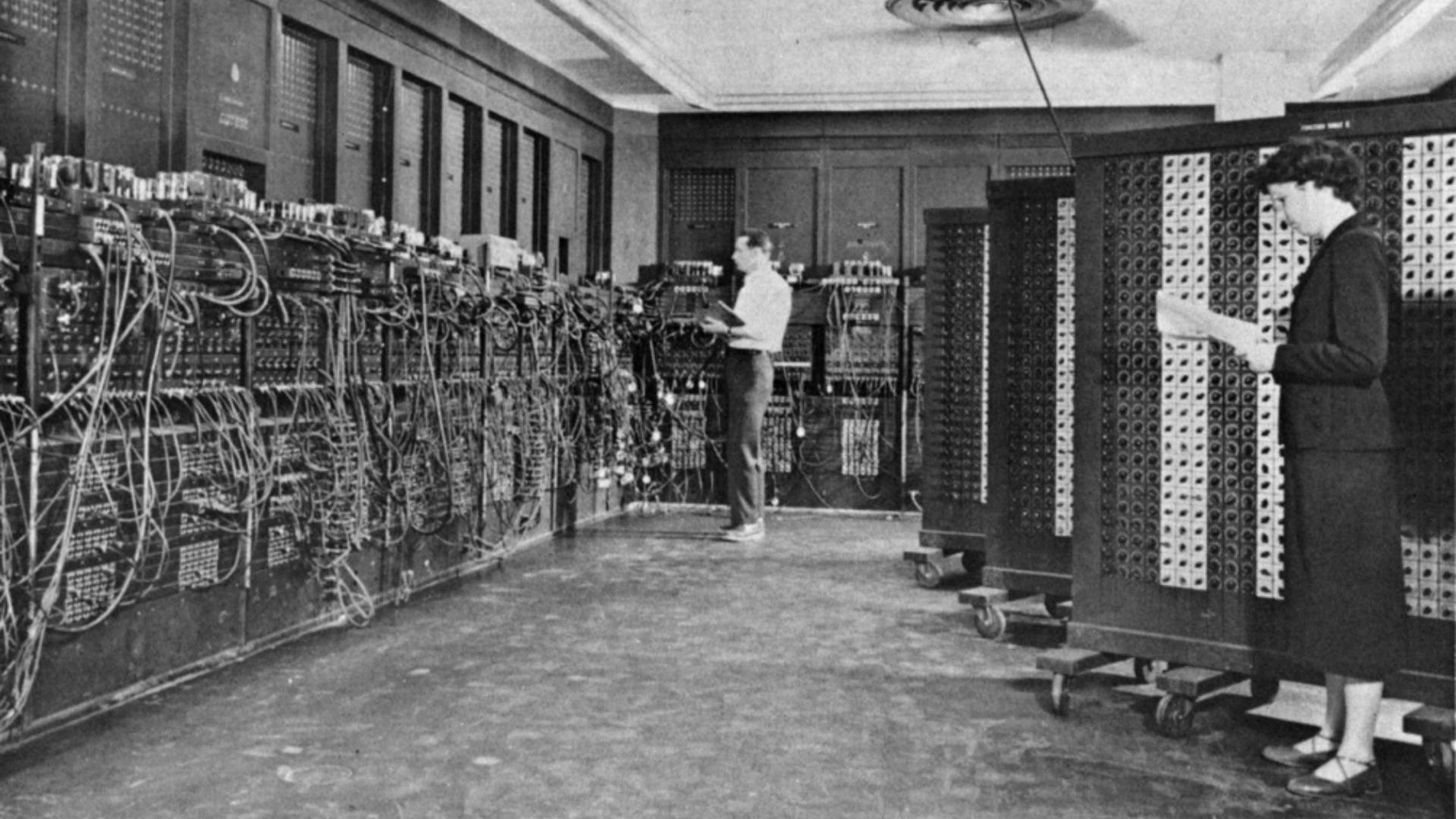
Prerequisite Knowledge

Basic familiarity with existing Big Data processing concepts/tools (Hadoop, Spark, etc.) is necessary. Familiarity with streaming concepts/tools (Samza+Kafka, Spark Streaming, Storm, etc.) is helpful. Familiarity with the Lambda Architecture is also useful.

Description

History has shown the limitations of existing streaming systems with respect to reliability, flexibility, and ease of use. The industry has responded in turn with the Lambda Architecture, a clever confederation of batch and streaming systems that provides low-latency, eventually-correct results, while maintaining the ability to respond to changes in upstream data. Lambda proponents have long argued that it's not possible to have all these things at once within a single streaming system. We respectfully disagree. :-)

We believe it is possible to build a streaming system you can rely on, making the Lambda Architecture unnecessary. In this talk, I'll cover:



LET'S TAKE A CLOSER LOOK





springone *2GX*

BATCH AND STREAM DEFINITIONS





LOOK AT COMMON USE CASES



CONCLUSIONS

DEFINITION: STREAM PROCESSING



**A system for processing an unbounded
set of data in an asynchronous manor.**

DEFINITION: BATCH PROCESSING





**DEPENDS ON
WHO YOU ASK**

“Batch applications run as special cases of stream processing applications”

- Apache Flink Documentation

**Batch processing is the processing of
a bounded data set without interruption
or interaction**

KEY DIFFERENCES



1

BOUNDED VS UNBOUNDED

2 SYNCHRONOUS VS ASYNCHRONOUS

3 STATEFUL VS STATELESS



STATE OF **STREAMING**



Google Cloud Dataflow



Tigon



STORM



samza





S4 *distributed stream computing platform*



samza





STATE OF **STREAMING**



EasyBatch



JSR-352

WHY THE END OF BATCH?





LATENCY

WHAT DOES THE END OF BATCH MEAN?





LIMIT
LATENCY



“ENTERPRISE GRADE”

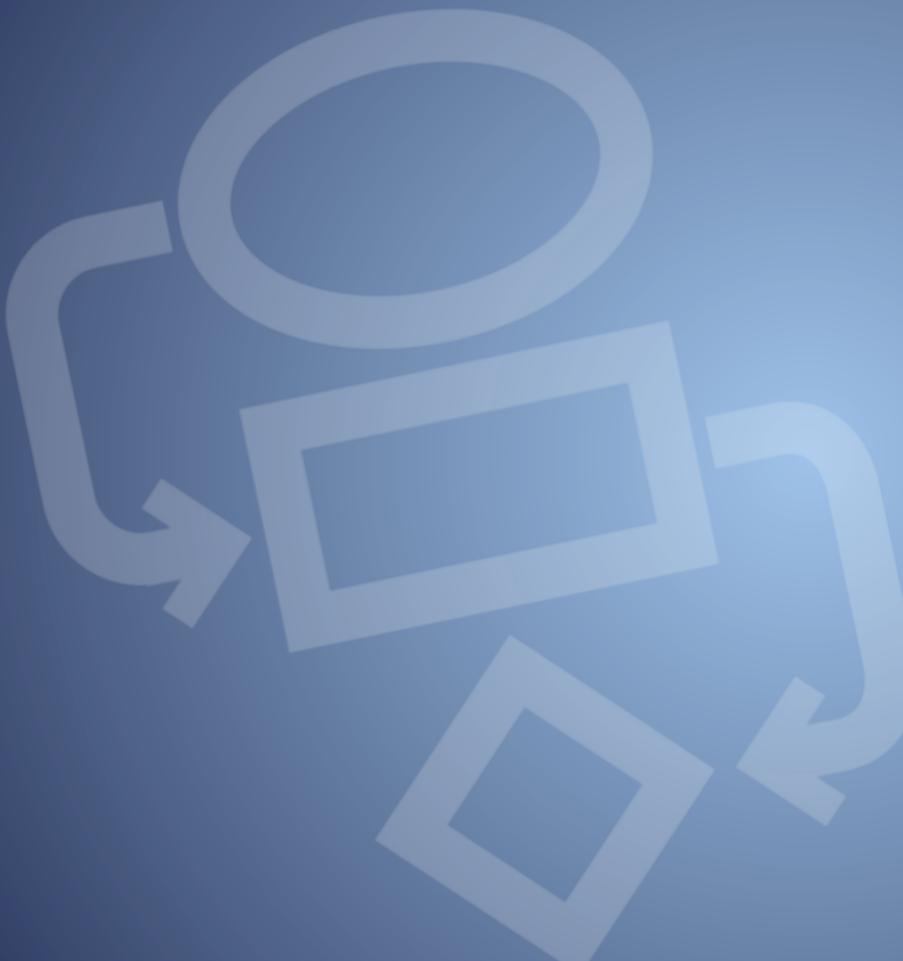


EasyBatch



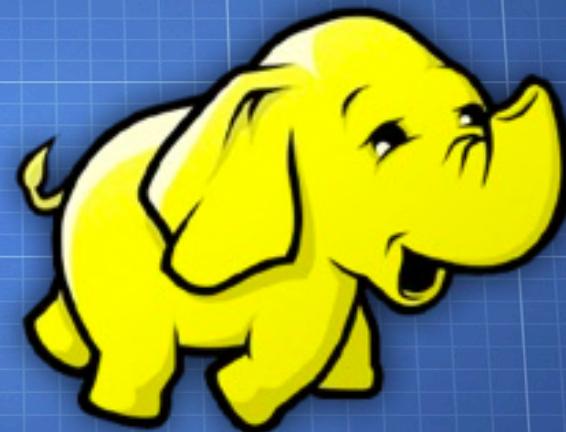


LOOK AT COMMON USE CASES



E.T.L.

DATABASE TO HDFS



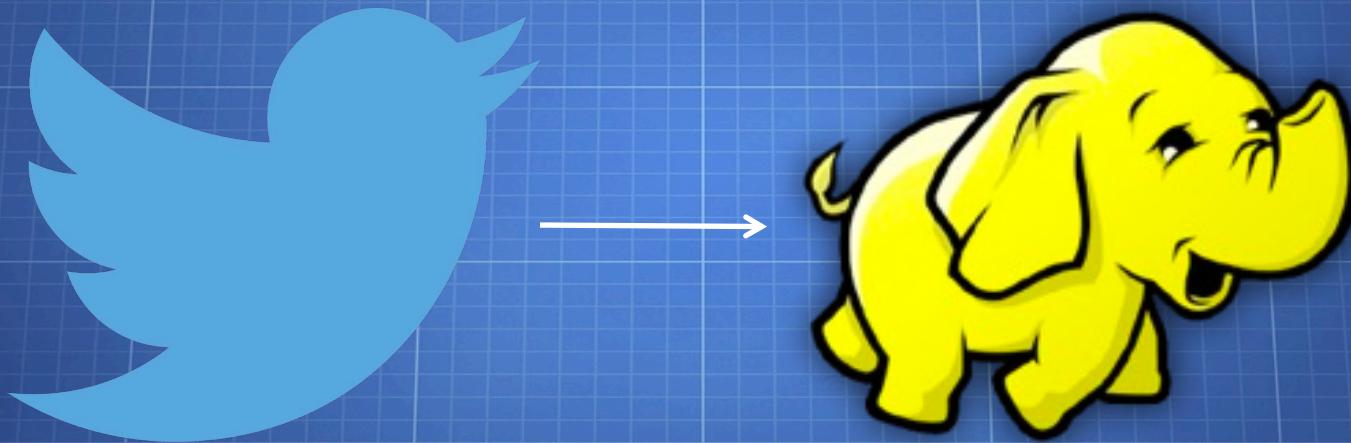


DEMO



STREAMING CAN BE USEFUL IN INGESTION USE CASES

TWITTER TO HDFS





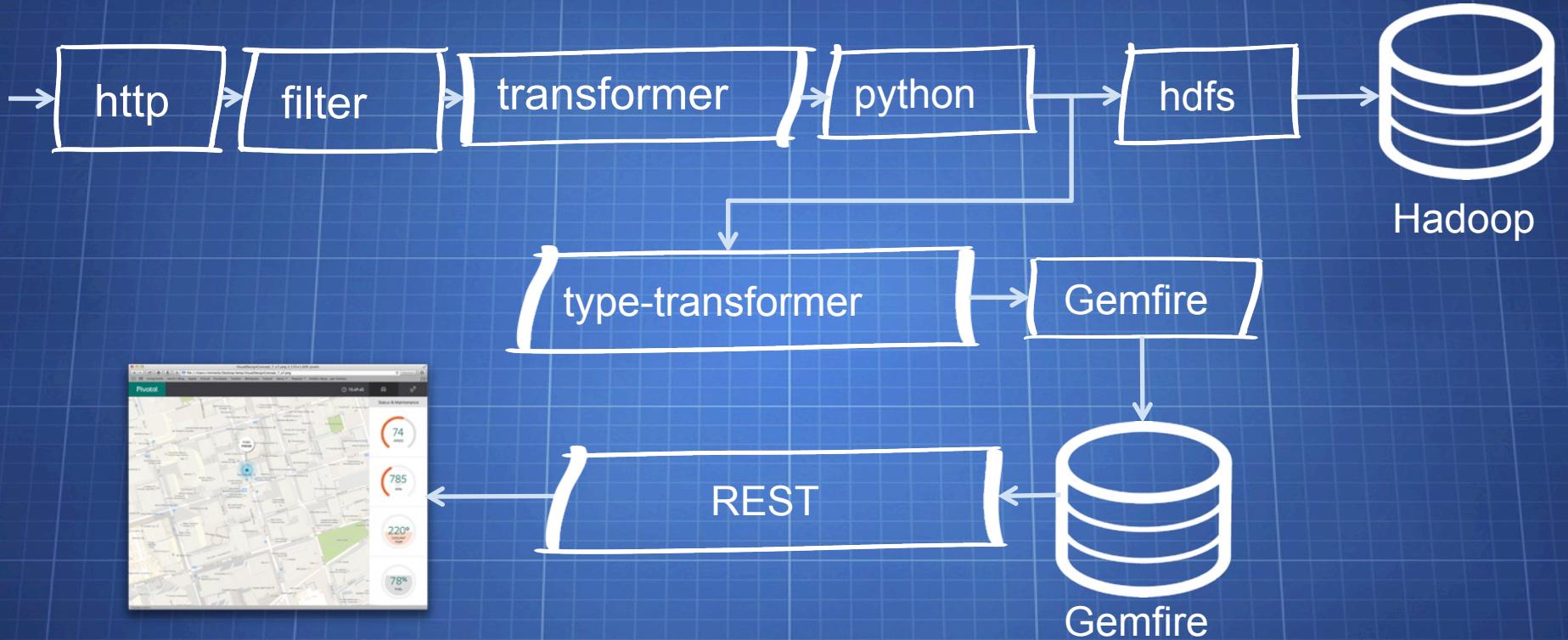
DATA SCIENCE

BATCH TRAINING PREDICTIVE MODELS





CONNECTED **CAR**





STREAMING
APPROXIMATIONS EXIST

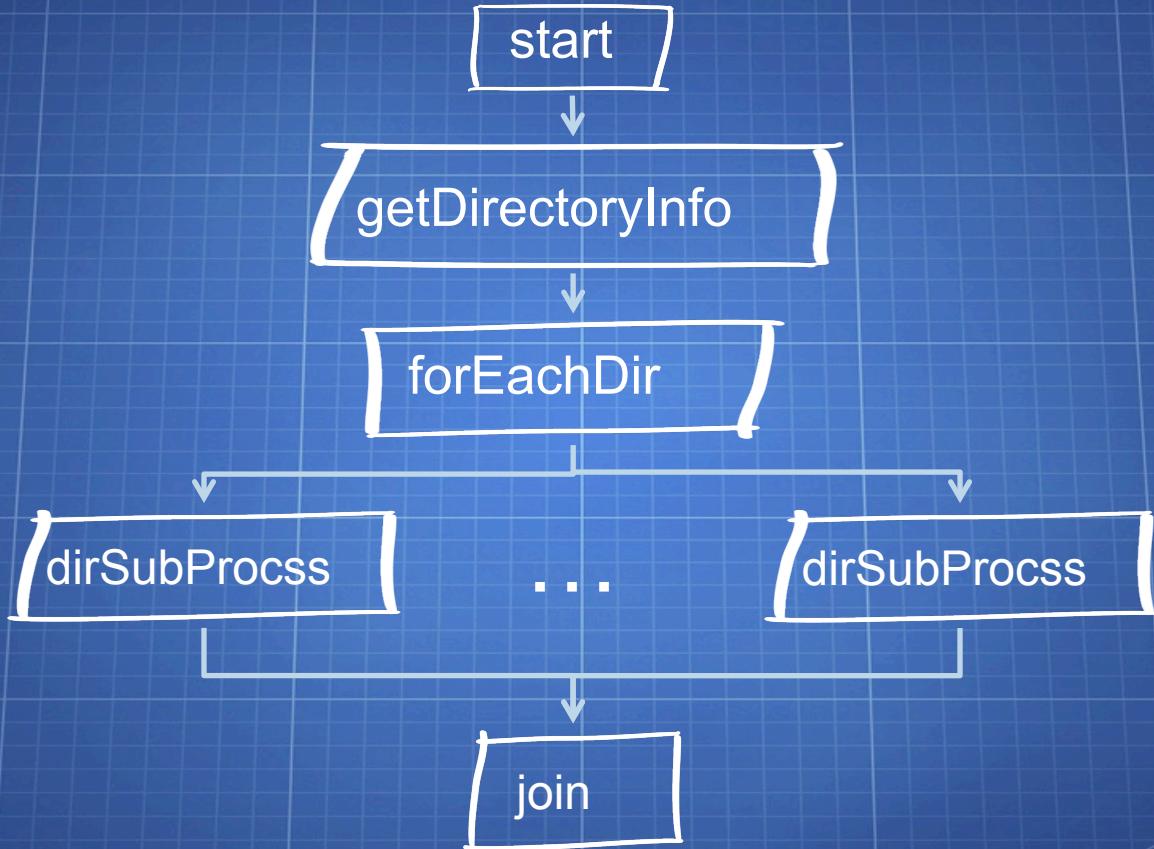


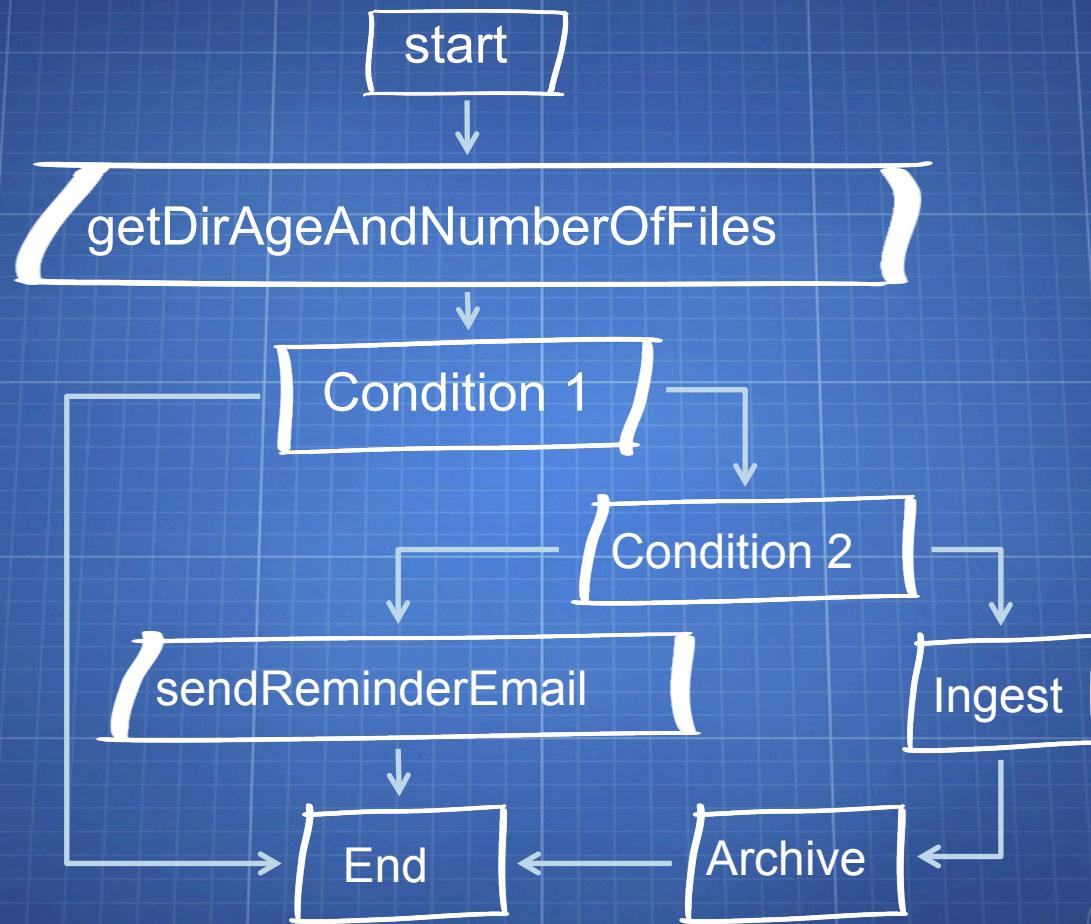
**BREADTH AND ACCURACY
DON'T MATCH BATCH**



WORKFLOW ORCHESTRATION

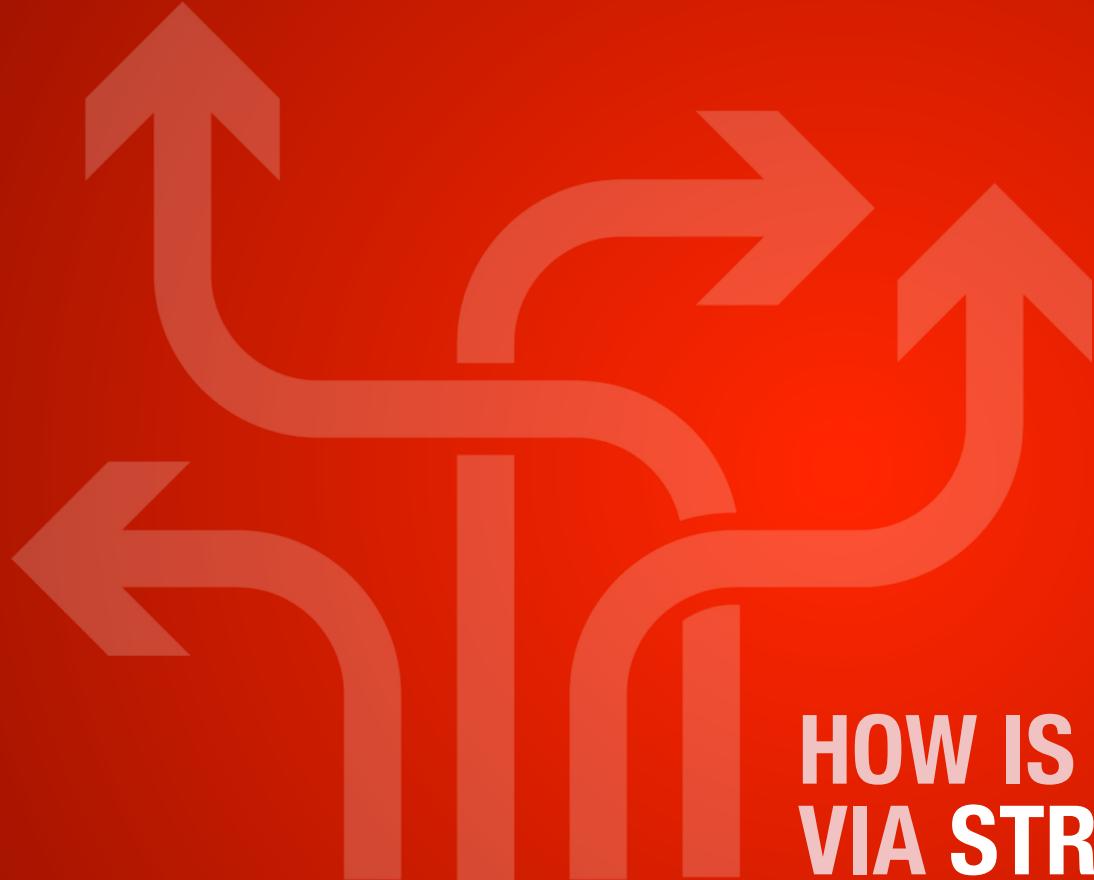








DEMO



**HOW IS THIS DONE
VIA STREAMING?**



NON-INTERACTIVE PROCESSING



**SAME AS ETL
PROCESSING**

PORT SCANNER





DEMO



**STREAMING
DOESN'T MAKE SENSE**

springone 2GX



**STREAMING
DOES MAKE SENSE**

springone 2GX



WHERE **LATENCY**
IS A **PRIORITY**



**DATA LOSS
IS ACCEPTABLE**

UNBOUNDED DATASET





**BUT DOES NOT
IN OTHER USE CASES**



HIGHLY COMPLEX



ERROR HANDLING NOT AS ROBUST



DATA GUARANTEES NOT AS ROBUST



WHERE BATCH IS BETTER



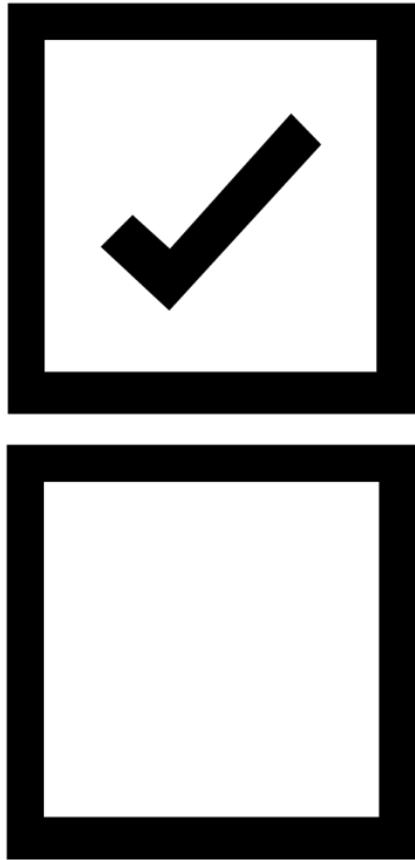
FINITE DATASET



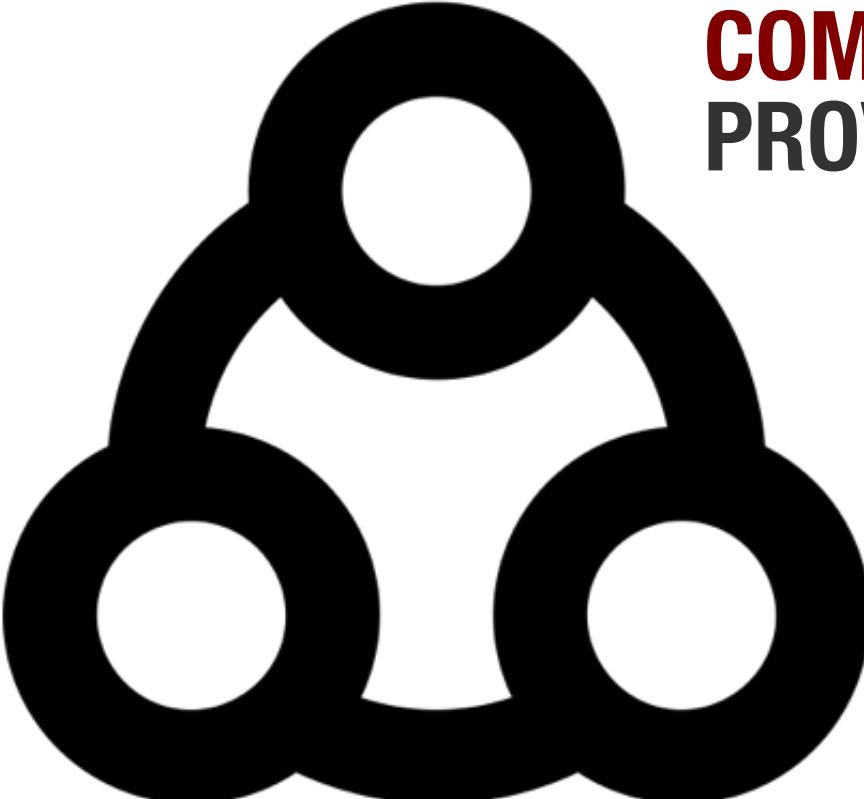
ROBUST ERROR HANDLING



RESOURCES ARE A CONCERN

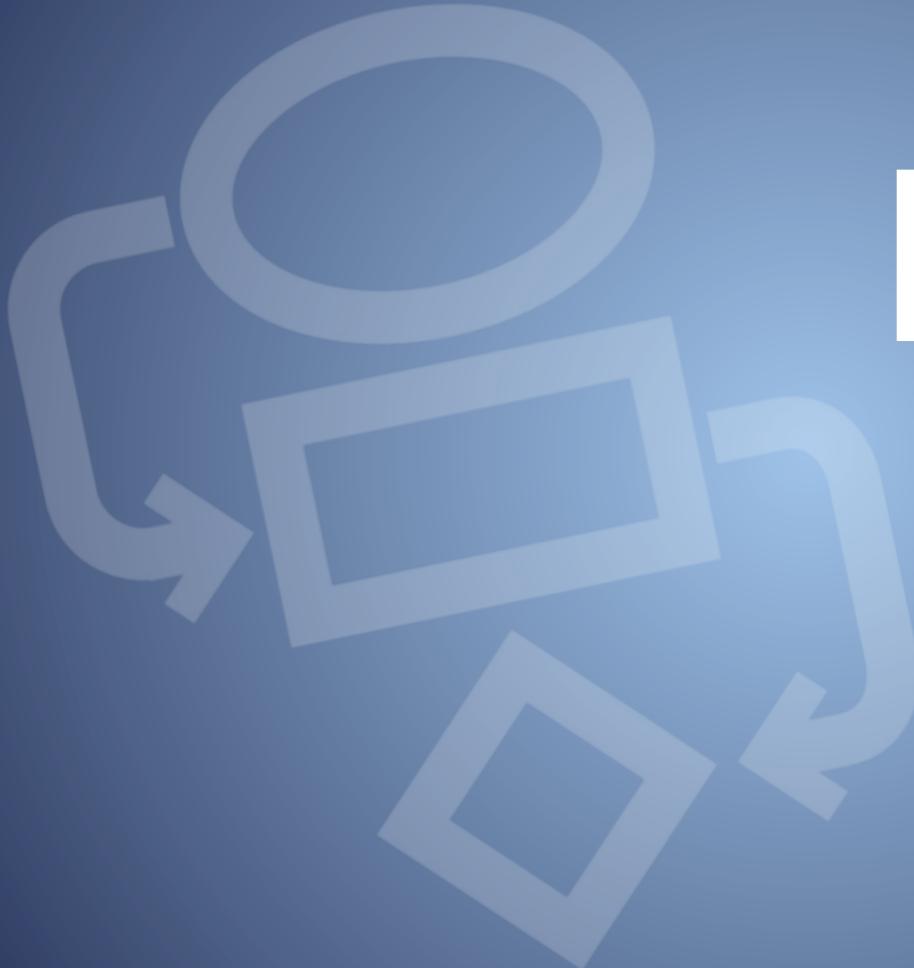


NOT EITHER OR



**COMBINING THE TWO
PROVIDES THE MOST POWER**

E.T.L.v2

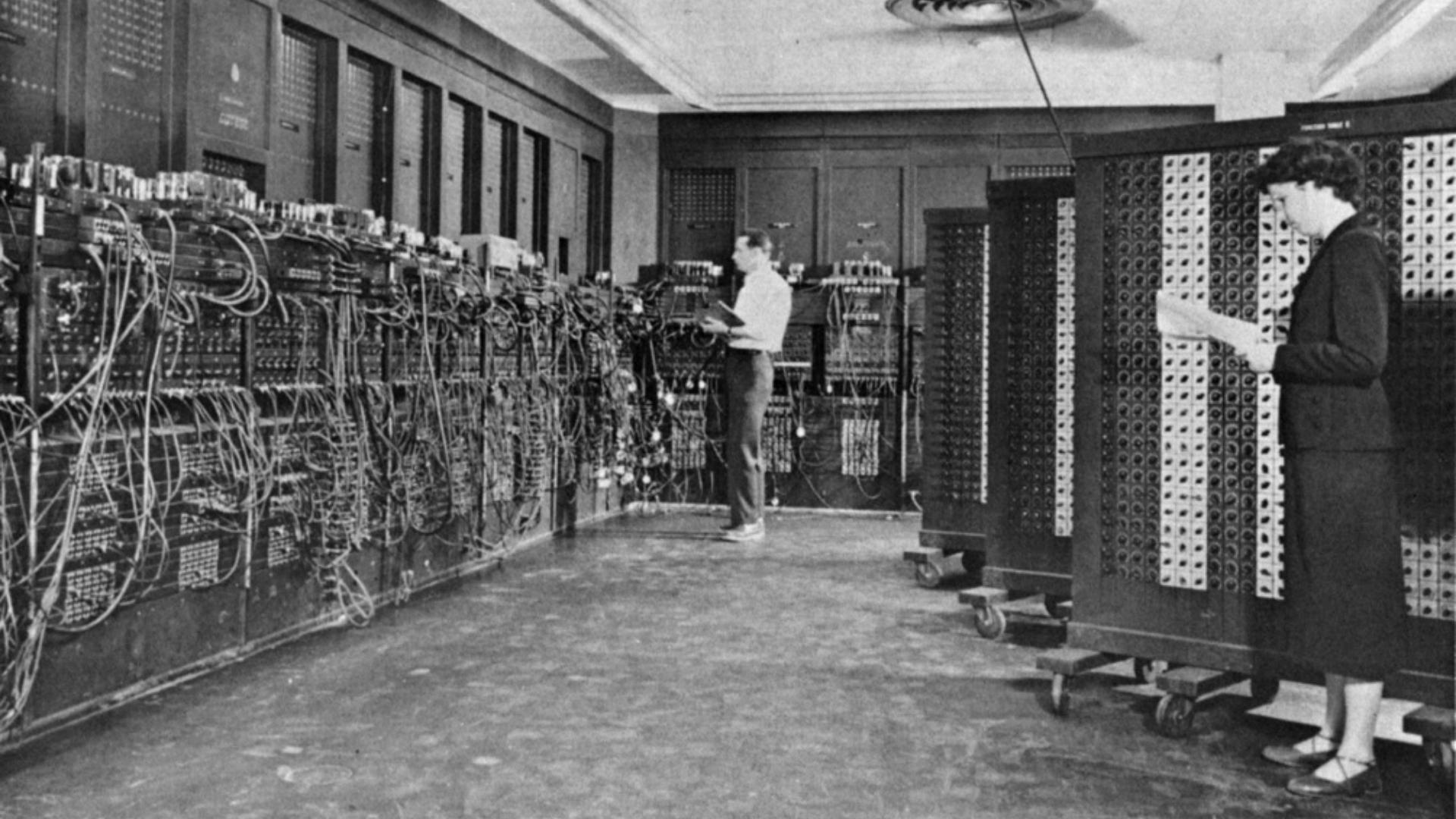


Database to HDFS v2

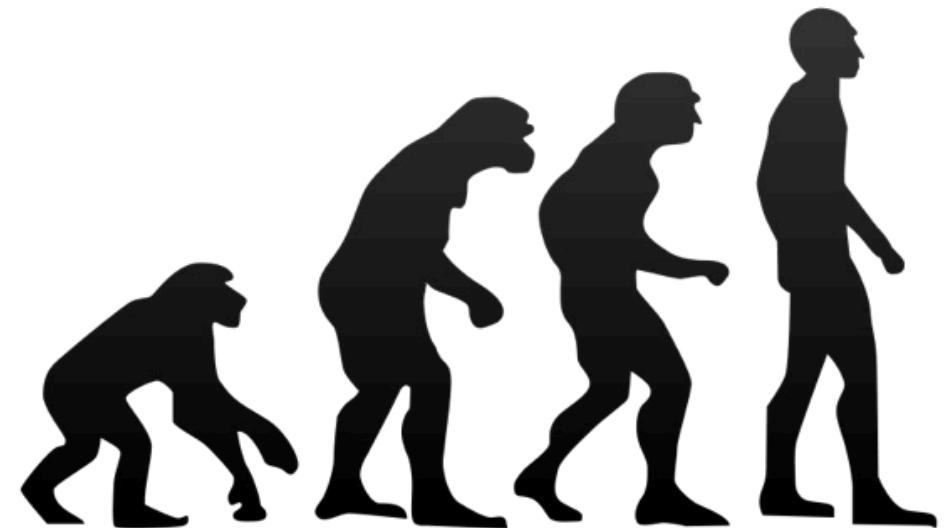




DEMO

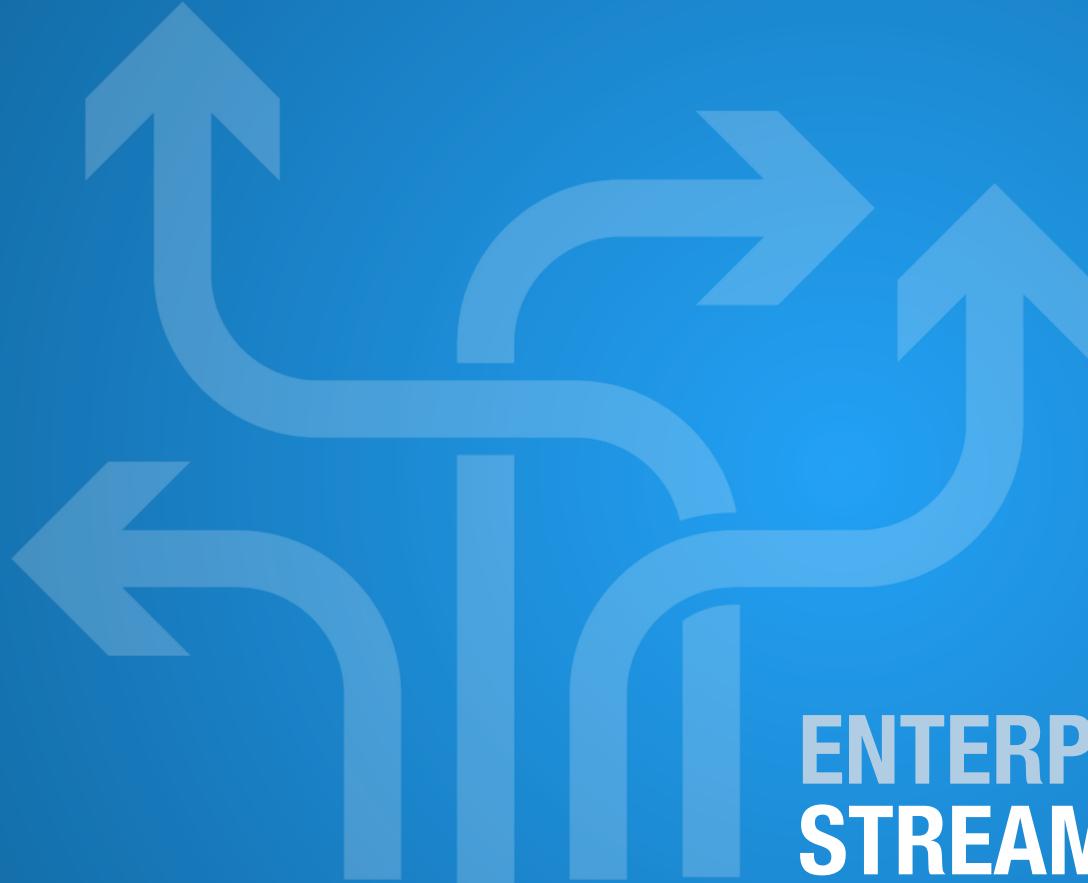


**BATCH HAS LASTED
BECAUSE IT MAKSE SENSE**



SPRING XD/DATA FLOW BEST OF BOTH WORLDS





ENTERPRISE GRADE STREAM PROCESSING

springone 2GX

BEST BATCH OPTION ON THE JVM





springone 2GX

SPRINGONE2GX

WASHINGTON, DC

Learn More. Stay Connected.



@springcentral



Spring.io/video

Check out Spring Cloud Data Flow on <https://spring.io>

Apache Spark for Big Data Processing – Salon N-P Up next!

High Performance Stream Processing – Salon N-P 8:30AM Tomorrow

Spring Integration Extensions Ecosystem – Here! 12:45 Tomorrow